Mime-Version: 1.0
Date: Fri, 5 Mar 1999 14:10:36 -0800
To: Harold_Varmus@nih.gov
From: "Patrick O. Brown" <pbrown@cmgm.stanford.edu>
Subject: Re: eprint
Cc: lipman@ncbi.nlm.nih.gov

Hi Harold,

I read your draft and thought it was tremendously exciting.

I have a few points of disagreement, the only really major one relating to the relationship of peer review to publication - I think it is very important that the peer review process not precede publication, but rather provide a way of stratifying and classifying papers that are published essentially as soon as they are submitted. This is a point worthy of serious discussion and debate, but I think there are compelling reasons for doing it this way.

I'm sending you a very rough and partial draft of a discussion of some of the key principles that I think should underlie the proposal, as well as some specific proposals. I think the specifics are important, to focus our discussion of more abstract issues, and to make sure that we can actually keep moving ahead on implementation.

I look forward to hearing your thoughts in return.

Best regards,

Pat

In case you can't open an RTF attachment, here's the text without formating. The attached version is easier to read.

_____
The Virtual Journal

(This is an absolutely central concept in the design of an electronic publication system, in my view).

None of us reads any journal from cover to cover, and no single journal provides any of us with all the published information we need. Yet each of us has always had our own individualized journal - a "virtual journal"... This virtual journal is the set of papers/fragments of papers /abstracts / titles/ news items/reviews/ letters/ advertisements...., that we each assemble on a continual basis, taking a couple of articles from the current

issue of Science, an article from a 4 year old issue of JBC, an ad from Biotechniques, a few letters from the NEJM/ a report from MMWR/ a news item >from Nature, a few paragraphs from Cell, a review from last year's Genes and Development, some abstracts from PubMed, a dozen journals' tables of contents.... We may read them individually on the spot as we find them, or this virtual journal may actually take on a kind of physical form as a folder-full of copied papers and excerpts of papers that we photocopy from the journals (or more often than not print them out from PDF files) and carry with us to read during a faculty meeting or on a plane...

The idea of the virtual journal not something new, idealistic and high-tech. It's just a description of the unacknowledged truth about how we use the scientific literature now.

The electronic "journal" that we are trying to create in place of the journals that exist now, is not a single journal, or some finite set of journals, but every individual's "virtual journal". To be more precise, we want to construct a system that makes the flow of information from the creators of published work to each of us as readers, through our "virtual journals" as ideal as possible.

The current journal system facilitates the process by which we create our virtual journals in one way - by grouping papers into discrete journals that (poorly) represent divisions by field, quality and importance, so that readers, based on their judgment of the likely density of work they care about in each journal (given their interest at the moment), can poke through them at different frequencies (mostly zero). The sorting system also helps to provide some markup of the quality of work, so that a literature search that yielded 40 possibly relevant articles could be prioritized for the laborious search through the stacks of the library, based on the weak but non-zero correlation between expected quality and the name of the journal. (Given 40 articles with the same title, by unfamiliar authors in an unfamiliar field, we would probably look first at the articles published in Science or Nature).

However, the current system greatly impedes the process of assembling our virtual journals by converting all the rich information collected in the peer review process into a meager few bits of information - the assignment to a journal. Moreover, by distributing the papers of interest to any given reader among a great many journals, each physically-separate, each with an expensive subscription price, each with a low density of papers that will wind up in any virtual journal, the current system makes the assembly of the virtual journal ridiculously, needlessly cumbersome and inefficient, and absurdly expensive (Stanford med school libraries spend $6,000/yr per full time faculty member, mostly for costs related to periodicals, and most faculty, students, and postdocs spend hundreds to thousands of dollars/yr., for journals of which only a small fraction will ever wind up in the "virtual" journal).

Arguments against the Virtual Journal model
An argument can be made that the process of browsing journals however inefficient, is valuable, and provides an opportunity for serendipity, inadvertant exposure to new ideas, and for meditative thought. I certainly agree with this argument. But all of this can easily, and much more effectively and inexpensively be recapitulated by means of the proposed eprint system.

An argument can be made that paper journals are a physically optimal medium for most conventional published work. Again I agree (though this is in part because conventional work has not yet had or taken advantages of the opportunities provided by electronic publication - for hyperlinking to cited publications, textbooks for background, background supporting data, audio clips of authors explaining their work, high-resolution images or video in addition to summaries of visual information, etc. etc.). Nevertheless paper journals are ideal in some ways. But the proposed system, by providing a single repository of the source material for each reader's virtual journal, can allow production of custom printed journals for each reader, or journals not individualized, but specially targeted to specific fields with the best, most attention worthy articles for each specialized field (overlapping sets between fields) - printed using existing technology. This is a real opportunity for libraries in the post-journal era - to devolop or provide the technology to provide in printed form the best approximation to the "virtual journals" for each member of the community they serve. There are commercial services that provide individualized, or small-run print jobs of this kind (even Kinko's does something very like this), at prices lower per page than most conventional journals, so this should be very manageable for most existing libaries. The customized printing of paper journals containing only papers that are likely to be read would result in a huge savings of paper, if they were to replace all our individual subscriptions to paper journals. Nevertheless, optimizing the reusability of the paper used for printing would be an important, secondary goal, which could be facilitated by a local (library centered) site for printing of individual paper journals.

Peer Review - its value to the author, the reviewer, and the reader:

For the authors, the value of peer review is: 1. to provide a final round of independent quality control (i.e. an astute and critical reader whose comments and suggestions help the author make the publication as good as possible) and, 2. to provide a kind of independent endorsement and recommendation of the published work to increase its visibility and impact.

For the potential audience - the "reader", it serves to classify and stratify the body of published work, by marking each publication as to

whether it might be of interest, given a reader's specification of his or her interests at any moment.

For the reviewer, the purpose is to contribute to the common good of the community by offering his or her time and expertise to help authors and readers; to communicate his or her own ideas and insights related to a publication under review; and to enhance his or her reputation, by building a record of constructive participation in this essential process.

The eprint system can take advantage of this convergence of the interests of authors, reviewers, and readers in the peer review process, by facilitating the connections between the involved parties, and by providing a controlled system for abstracting key information provided by peer review to help guide readers in their search for information.

How should the peer review system operate?

The process begins when an author decides to submit a "work" for publication, by sending it to the eprint site. The only step between submission and public posting of a newly submitted work at the eprint site is a cursory review by a member of a large voluntary group. (I suggest that we assign this role on a rotating basis to members of a very large group of scientists, such as all PI's on NIH, NSF, etc. grants). Each newly submitted paper is immediately put in a holding area, made visible only to a group of say 100 scientists (out of a pool of tens of thousands) that changes daily. Members "on call" on a given day are asked to look over as many papers in this site as they can, and to "pass" any that meet minimal standards. "Passed" documents, say those which are judged to meet minimal standards by two of the screening panel, are immediately posted at the publicly-accessible eprint site.

At this critical step - the decision to publish, there is no judgment of the science -submissions are just checked to exclude invective, commercial advertising, gibberish and completely inappropriate language. Abject stupidity is acceptable - the author takes the risk of exposing him/herself, but the reading audience is at negligible risk of wasting any extra time as a result, since publications will be classified by the subsequent peer review process so that the real junk can be filtered out unless a reader deliberately chooses to read it (no doubt someone will put up a third party site entirely devoted to the most stupid, worthless or crazy material submitted to the site). The really bad papers will no more pollute the literature than they do now - there are thousands of junk journals that none of us reads, and they are of essentially no consequence to us (except as a financial drain) and at least with this system, they won't take up space in library shelves, or run up subscription costs.

Harold: this is a significant point of difference between what I would propose and your draft proposal. >> As soon as the work is posted at the public eprint site it is, by any sensible definition, published. The publication status of the work doesn't change through the review process (except in that a revised version might be substituted for the original version, with the original version remaining as a record linkable from the new version). It is published the instant it is "passed" for posting at the site, and it doesn't later move to an "officially published" status based on the reviews. Rather than determining whether and after what delay a paper is published, what the reviews provide is: 1. a basis for classification of the work according to how much attention it warrants, and >from whom; 2. a source of critical feedback to the authors that may lead to substitution of a strengthened version of the publication; and 3. Critical commentary that can highlight important contributions or deficiencies in the publication, for the benefit of readers.

The work may, at the moment it is posted, already be accompanied by formal reviews provided by qualified reviewers whose reviews were solicited by the authors, or alternatively, peer reviews may be solicited by the authors at the time the work is initially posted. The solicitation process can be facilitated by the eprint system, in at least 2 ways, as follows:

1. One opportunity flows directly from a critical positive feature of the electronic preprint system. Most manuscripts with any value at all, however specialized, will be read by individuals in their field as unreviewed preprints - this is the same as going to a seminar, a symposium, a poster session, or reading a preprint - people look at unreviewed work in their field all the time - selectively. The people who would look at a given paper in preprint form will tend to be the people who are most qualified to provide a critical and thoughtful review, the same people who would be likely to be solicited for a formal review in the current system. So we will ask, as a matter of "good citizenship" - which is the principle motivation for reviewing papers now - that people who read a paper, particularly one that has fewer than two reviews already posted, provide a critical review of the manuscript. The first several times a publications is downloaded, (or until adequate reviews are posted), the reader will be asked to volunteer to provide a formal review (which will need to be submitted via a qualified reviewer, as designated by an "editorial board", see below). We can try to establish this as kind of a standard of good scientific citizenship.

2. The author designates the fields to which he or she believes the work relates using a controlled definition of the field. On the basis of this designation, and perhaps suggested reviewers identified by the author, qualified reviewers from a list of reviewers assembled by the "editorial board" (see below), can automatically be contacted (in two ways - by email, and by a notice that will appear when they log on to Pubmed or to the

eprint site). The reviewers will be asked ("yes or no") to provide a review - "you have been suggested as a reviewer of "title" by "author" URL####" please consider providing a review of this publication. Please indicate whether you agree to do so". Once a prospective reviewer agrees to submit a review, he or she will automatically be reminded periodically until the review is posted. Very tardy reviews can be cause for a prompt >from the "editorial board".

A feature important in making the process collegial and constructive rather than adversarial and invidious is that reviewers will be asked to communicate their comments to the authors directly and privately, before posting a formal review, to allow the authors to modify the publication, if they wish. This direct communication between reviewer and the authors can be optional, but should be the expected norm. Once the reviewer and authors agree on revisions, or agree to disagree, the review is posted, accessible as a link from the original or revised manuscript (or directly, if one wanted to search directly for reviews, eg., by a specific colleague).

The reviewers can provide a free form review like the typical written review of a paper for a journal, but will also be expected to classify the work using a more controlled format (related to the systems used, eg., by Science and PNAS for rating papers on a scale). We might require that at least 2 independent reviews will need to be posted in order for a publication to be classified as described below. Additional reviews submitted by qualified reviewers at any time after publication will also be weighed, so that a publication's visibility can rise or decline at any time (even years) after its initial publication, based on the ongoing assessment of its value. It will be natural for these reviews to be posted even long after a work is published, by readers who have a strong opinion that a publication was under- or over-appreciated when first reviewed.

All reviews will be signed by the reviewer and accessible in full as links >from the reviewed publication.
Having the formal peer reviews signed and published will make the process more constructive, creative, and rewarding for the reviewer. The peer review becomes an opportunity for the reviewer to make an original, insightful suggestion that becomes part of the permanent record and can even be cited if it is relevant to the citation of a paper, or if it is an important original contribution in its own right. The published review provides part of a permanent public record of the reviewer's contributions to the integrity and collegiality of the scientific process. Indeed, the rigor, compassion, constructiveness, creativity and volume of a scientists complete body of work as a peer reviewer can become a useful criterion for judging what kind of a colleague a prospective hire might be, and for evaluating scientists for promotions, etc.

A specific proposal for structuring the classification and stratification of published work based on peer review (intended to capture the useful features of this process as it now occurs through division of the literature into journals of various fields, levels of specialization, and reputation)

We start by devising a hierarchical classification of fields or topics, which NCBI can develop based on the implicit "field" structure of the current literature indexed by medline. First, essentially every specialized field or topic represented by at least 10 publications per year is identified (fields may be defined based on keywords and title words used in medline indexed publications, for example; and by the conventions embodied in specialized journals and scientific societies, or by collecting data on the words used in searches of Medline). The Fields are not mutually-exclusive, but can overlap. Fields are then assigned "levels" that reflect their breadth and their "volume" (rate of publications in the field) as follows: A level 1 field is one that encompasses more than 100,000 medline-indexed publications per year (eg. Biochemistry, genetics, medicine). A level 2 field is one that encompasses 10,000 - 100,000 publications per year (eg., virology, neuroscience, pediatrics); A level 3 field is one that encompasses between 1000 - 10000 publications per year (eg. AIDS, DNA enzymology, retrovirology) and so on to level 6 (less than 10 per year). Obviously the scale could be continuous rather than discrete [eg. the level could be defined as log(10) of the number of papers per year]. NCBI should be able to classify fields on this scale using some kind of abstraction from the last several years of Medline, and revise it on an ongoing basis.

Peer reviewers submit their reviews through a web interface. They are asked to identify the fields to which the publication under review relates. They are prompted based on their responses to choose a field in the formal hierarchy described above. They are specifically asked in turn to identify the broadest field in which the publication would be among the 10 most noteworthy of the year, then the fields in which it would be among the 100 most noteworthy of the year, and then any fields in which it would be among the 1000 most noteworthy of the year (this process can be helped by automatic prompting with fields one level above and one level below their suggested field in the hierarchy). Some publications would not fall into any of these classifications. If this is due to a deficiency of the classification system, the reviewer could contact the administrators of the system or the corresponding editor and ask for an addition or change to the field classification system. If this is due to the work being essentially worthless, then the reviewer would simply enter no field for any of these levels of noteworthiness. If a reviewer indicates that a publication is in the top 10 of a level 3 or higher field, the top 100 of a level 2 or higher field, or the top 1000 of a level 1 field, then he or she is asked to provide a justification for that prestigious designation. A typical publication may fall in more than one specialized field, and be assigned a

high prominence in one field, and a lower prominence in another.

The resulting classification provides a natural way to organize and stratify the publications, matching what the readers want to use in compiling their virtual journals. Each publication is given a numerical index of its "attention-worthiness" in each of several fields at several nested levels of specialization. Given any reader's specification of a field of interest, the publications in that field could be provided selectively on the basis of timeliness and the peer reviewers' ratings of attention-worthiness. If one wanted to replicate in an idealized form what people imagine they get from the existing system of journals, one could simply take a quarter of the biology (level 1) top 1000, and 2 of the top 10 papers in each level 2 field, and put them in "eScience", and put a similar mix in "eNature". Half of the top 1000 publications in Virology (level 2) could be published in Journal of Virology, and so on.

A concern about giving up the journal system is how young scientists would be judged for hiring and promotion, if we were to abandon the current system of delegating their evaluation to an arbitrary collection of anonymous reviewers and editors. Although I think there are many reasons that this concern is specious, one answer is that we could reconstruct an analogous, but better system, for example: When citing a publication in their CV, authors might, in lieu of the journal name, mark a publication as "class 2.1 in Virology, class 4.1 in Biochemistry...." (where the class here could be the parameter representing the median percentile ranking of a paper in the corresponding field, as assessed by the peer reviewer - eg. if two reviewers rate a paper as in the top 100 per year in a field with 10,000 publications per year, this might be a class 3 paper).

How can we provide a ready source of willing reviewers for authors who want to forgo the author-selected reviewer mechanism, or to supplement reviews provided by the author? Why not ask highly reputable groups, with some insight into the performance of their colleagues as potential reviewers (based for example on their experience as editors), eg., NAS, AAAS, HHMI, other societies with high standards, editorial board members of scholarly society journals - to provide lists of qualified reviewers as a function of fields, for a set of fields (as defined above). Reviewers chosen from these lists would be automatically notified when a publication is posted in their respective fields.

We can also put together a broad, but still highly selected group using many of these same sources, numbering, say 1000-2000 scientists, to serve as a kind of editorial board - overseeing the review process - checking on the quality of reviews, soliciting reviews personally in some cases, monitoring the performance of reviewers so that they can be commended or perhaps removed from the solicited review list, adding new reviewers to the list of eligible reviewers, based on the quality of their unsolicited or tentatively-solicited reviews, or on their newly-recognized stature in a

field, etc. We can organize this editorial board, perhaps based on their own assessment of the fields in which they are qualified and willing to serve as "editors". Their function would be unlike editors in the conventional journals in the sense that they would never make a decision on whether a work ought to be published, only to provide supervision of the review process to make sure it performed adequately.

An important additional mechanism for organizing/stratifying and adding value to the work published at the e-biomed site will be "third party" web sites (or printed journals), established by individuals, societies or for-profit publishers, unaffiliated with "e-biomed". These entities would not play a role in the primary publication process, but would serve as intermediaries in providing selected and enhanced collections of published work to their readers. They would have an incentive not only to watch new submissions that are obviously in the field of interest, but to search and discover new or old work that would not readily have been found or appreciated by their audience, but whose value can be highlighted by the third-party site (added value).

Making the most of the electronic medium/internet distribution:

Authors should be encouraged, and helped as much as possible, to make their papers as accessible as possible to its vast potential audience, and to maximize their use of the potential of the medium. Trivially, we should encourage authors to provide expanded abstracts to facilitate the browsing/searching process. We should encourage use (but not gratuitous use) of video, 3D images, audio, and rich visual display of results or models. We should also strongly encourage authors to provide, as part of their publication, a parallel version of their work intended to be understood by a non-specialized audience (even a lay audience), exploiting the fact that the potential audience is anyone with a internet-connected computer.

International access:
One important and sensitive issue is the "internationalism" of the site. I think we could consider the possibility of having scientists in other countries volunteer to help in the screening and classification of submissions in languages other than English, but in the short term, we'll need to restrict the site to English language papers. The question of access is still one of the strongest arguments in favor of the eprint mechanism. The less affluent countries will benefit the most - the cost of a computer suitable for viewing and printing out material from the internet is less than the subscription cost for a single average journal. Most of those less affluent countries have multiple disadvantages - limited access to journals, delayed receipt of the journals, even more limited access to

archived back issues of journals, the language barrier (most of the best journals are written only in English - a serious obstacle even in Japan), and the fact that in the more priviledged institutions the most important scientific news is usually old news by the time it's published. Electronic, rapid distribution of virtually all scientific reports (combined with conversion of the archival literature to electronic form) would even the playing field in all these areas. Any individual or library with an internet-connected computer would have access to the same publications as the most favored institutions and individuals. There would be much less delay between the spread of information by the "grapevine" to which only the fortunate are connected, and universal distribution of important new results and ideas. And although they are far from perfect, language translation programs could be used to convert the electronic text >from English (or Japanese, or whatever), into the native language of the user. I don't think that there will be any slowing in the trend toward English as the universal language of science, but it would be a good thing if we helped make this fantastic body of information accessible to people with little or no English (eg. College or High school students in other countries, who might thereby be drawn into scientific careers, and learn English in the process).

Wide public access:

With the proposed system, every library can have the same collection of periodicals as the best university. Institutions off the beaten track will no longer be so late in hearing about important new ideas and results that are old news by the time they are published, to those of us lucky to be at the most favored institutions. Students at all levels and the public can watch the scientific process, with all its debate and uncertainty, rather than getting only the dumbed down textbook version, with its false implication of certainty and finality. There will be a real opportunity to build a teaching level into the system that will use the universal access to raw scientific results, and discussion as a great teaching resource, and to promote healthy scepticism and curiousity about science and medicine.

Why NIH would be crazy not to invest in converting back issues of print journals to electronic form, and obtaining copyrights for free universal access.

Following the launching of an electronic publication system, NIH should give high priority to the conversion of the printed medical literature to electronic form.

Almost every day, almost every one of us wants to look up some information published in the past, perhaps cited in a current article, or found in a

Pubmed search, and to quickly take a look at it - sometimes in detail, sometimes at a minimal level of depth and rigor. Today, the rate limiting step (often means it doesn't happen at all, or you stop at the level of reading titles or abstracts), is going from the citation, or from a title or abstract in pubmed to the full text, or a selected nugget of text. This step usually involves a trip to the library, followed by searching through bound volumes. At best this is rate limiting, more often, it is an energy barrier that aborts this little intellectual exploration. This energy barrier will become even more obviously limiting when current publications are all on line, with hyperlinks to citations. Those citations that are not available electronically as full text will be almost infinitely less accessible and useful than those available electronically and their value will be greatly diminished. And yet, the knowledge and information contained in those old publications still has plenty of potential value. Almost every day, all of us still rely on information found in articles published 2, 5, 10, 20 years ago. They are read at a much lower density than contemporary articles, but are still indispensible.

Their economic value to the owners of their copyrights is negligible, due to the low current demand for fresh copies, and the low density of demand to support pay per view access.

The cost per page for conversion of text and figures in a journal article to digital text and figures with SGML markup is estimated by David L to be $2.00 . If we estimate that all the publications in the biomedical sciences for the past 50 years add up to 100 million pages (a major overestimate, since the current annual production is probably less then 4,000,000 - based on 400,000 articles indexed by medline in 1998, estimating 10 pages per article, and assuming the average output for the past 50 years is half of this output); then we get a total cost of 200 million dollars for digitizing this entire body of literature. This is surely much less than the average annual cost that NIH spends for journal subscriptions alone. But consider instead the cost in time. Suppose that there are 100,000 biomedical scientists in the US, and each spends 5 minutes a day that would be saved by electronic access to journals that currently require a trip to the library. This is 1,000 minutes/year times 100,000 individuals. If we value that time at $30/hour, then the lost productivity costs $50,000,000 per year. So the investment in digitizing journals will clearly pay off quickly. But the greater value is in the recovered value of all the virtually inaccessible information that would suddenly be at our disposal. And not only at our disposal, but at the disposal of everyone in the world who can access the internet (still not available enough to poor countries and communities, but much more so than journals in university libraries).

A concrete proposal: We probably don't sacrifice much by triaging the journals to be digitized, recognizing that a few dozen are critical, a few hundred worthwhile, and the rest perhaps not worth the trouble at all.

Suppose the top 500 journals can be identified, and that they average 1000 pages each per year. This is 500,000 pages total. We start by converting all of these journals dating back 30 years (15,000,000 pages) to electronic form in the first year. This will cost between $7.5 million and 30 million. If the journals charge for the copyrights, the cost might be increased by a further million or two. The "top" journals could be identified somewhat objectively based on some combination of their citation index, their paid circulation, and the frequency with which abstracts from those journals were called up by users of Medline (A couple of months of tallying this information by NCBI should give a pretty good index). One might want to rank the journals and years and proceed progressively from the most heavily used on down the list. My guess is that this part of the project could be done in a couple of years. It might even be possible to get the Bill Gates foundation, or some other foundation to finance it if NIH couldn't.


eprint.3-4.RTF

Patrick O. Brown
Howard Hughes Medical Institute
& Department of Biochemistry
Stanford University School of Medicine
Stanford, CA 94305-5428

Tel: (650) 723-0005
Fax: (650) 723-1399

http://cmgm.stanford.edu/pbrown